Proceedings of the Society for Computation in Linguistics

Volume 3 Article 2

2020

The stability of segmental properties across genre and corpus types in low-resource languages

Uriel Cohen Priva
Brown University, uriel_cohen_priva@brown.edu

Shiying Yang
Brown University, shiying_yang@brown.edu

Emily Strand
Brown University, emily_strand@brown.edu

The stability of segmental properties across genre and corpus types in low-resource languages

Uriel Cohen Priva* and **Shiying Yang*** and **Emily Strand**
Brown University
Department of Cognitive, Linguistic, and Psychological Sciences
190 Thayer St., Providence, RI 02912, USA
uriel_cohen_priva,shiying_yang,emily_strand@brown.edu

Abstract

Are written corpora useful for phonological research? Word frequency lists for low-resource languages have become ubiquitous in recent years (Scannell, 2007). For many languages there is direct correspondence between their written forms and their alphabets, but it is not clear whether written corpora can adequately represent language use. We use 15 low-resource languages and compare several information-theoretic properties across three corpus types. We show that despite differences in origin and genre, estimates in one corpus are highly correlated with estimates in other corpora.

1 Introduction

One of the challenges facing corpus research in phonology is the absence of detailed cross-linguistic phonological corpora. When a phonological trend is found in a language or a language family, e.g. OCP in Semitic (McCarthy, 1986), does it extend to other languages too? Variation-friendly versions of Optimality Theory (e.g. Anttila, 1997; Boersma, 1998; Goldwater and Johnson, 2003) predict that obligatory constraints in one language would appear as trends in other languages too, e.g. languages without grammatical final devoicing should have fewer voiced codas than voiced onsets. This rigor is difficult to achieve without detailed phonemic lexicons.

The Crúbadán corpus (Scannell, 2007; cf. Zuraw, 2006) provides word frequency files for thousands of languages, often based on Bible translations and Wikipedia. The Linguistic Data Consortium (LDC) has provided data for many languages in various formats, e.g. conversation transcripts and newswire, from which word frequency files could be easily generated (for a few languages, LDC provides such data directly). An intriguing new source for word frequencies is the Open Subtitles Corpus (Tiedemann, 2009), which collects subtitle data for multiple languages. Therefore, it potentially represents spoken language better than Bible translations or Wikipedia.

There are several challenges in using word lists for research in phonology. First and most obviously, some procedure needs to be applied to translate alphabetic representations to phonemic representations, if such a procedure is possible.[1] But even in cases in which a clear correspondence between the alphabet of a language and its phonemic representation does exist, we may suspect that the data itself is inadequate, or not representative of the phonemic trends of the language. For instance, Daland (2013) discusses *burstiness*, or the possibility that otherwise low-frequency words could bias a sample due to them being over represented in a particular subset of the corpus. A good example of this effect can be found in the Crúbadán entry for Indonesian, in which the word *Indonesia* is the 14th most frequent. This is due to the fact that the word frequencies were created from the Indonesian Wikipedia, a corpus in which the word *Indonesia* is very frequent. For comparison, the word *Indonesia* is not among the 1,000 most frequent words in the word frequency files derived

*Corresponding authors

[1] For some questions, using the alphabet directly may be enough (e.g. Piantadosi et al., 2011), but for phonological questions, the use of the alphabet as a proxy for phonemic representations is suspect.

Proceedings of the Society for Computation in Linguistics (SCiL) 2020, pages 1-9.
New Orleans, Louisiana, January 2-5, 2020

from an Indonesian newspaper collected for Cohen Priva (2017).

Despite burstiness, recent findings suggest that segment frequency, predictability, and informativity values converge to their model values rather quickly (Cohen Priva and Jaeger, 2018), which may follow from the segmental domain being considerably more dense than the word-and-above domain. However, their findings compared subsamples of a corpus to the entire corpus, rather than different corpora to one another. Furthermore, word frequencies were established using spoken corpora. Would it be valid for other studies to rely on word frequency lists from different genres, often less representative of spoken language? An additional limitation is that their findings were based on only one language with millions of word tokens in the entire corpus (the samples were substantially smaller). Our goal in this paper is to assess whether similar findings arise without these limitations, e.g. would Crúbadán-based data be similar to spoken data from the same language, using smaller corpora, and many different languages.

2 Methods and materials

2.1 Word frequency lists

We used word frequencies from three corpora, the Crúbadán Corpus (Scannell, 2007), the Open Subtitles Corpus (Tiedemann, 2009), and conversation transcripts (some of them scripted) from the IARPA Babel program (Adams et al., 2017; Andresen et al., 2019, 2018, 2017; Andrus et al., 2017b; Benowitz et al., 2019; Bills et al., 2015, 2018, 2016; Conners et al., 2016). We only used languages that appeared in the Open Subtitles corpus, or were part of the IARPA Babel program. For every language, we ranked word type by token frequency, only considering words that had the same or more occurrences than the 30,000th ranked word. Additionally, we excluded words that our rules could not translate as well as words whose frequencies in that corpus were lower than 5. Furthermore, we did not use Georgian from the Open Subtitles corpus because we determined that although the words consisted of Georgian script, many were not actually in Georgian, but possibly in Russian.[2] We similarly excluded Haitian Creole from IARPA Babel (Andrus et al., 2017a) because the spelling convention was not consistent with written Haitian Creole. We also excluded words that had any uppercase letters in them in order to discard of irrelevant data, including but not limited to names, acronyms, and companies. The resulting number of types and tokens per corpus are listed in Table 1 for Open Subtitles, and Table 2 for IARPA Babel.

Table 1: Open Subtitles vs. Crúbadán type and token frequencies

Language	Open S. types	Open S. tokens	Crúbadán types	Crúbadán tokens
Bulgarian	23,100	342,000,000	21,300	1,160,000
Catalan	17,700	2,790,000	17,900	1,510,000
Greek	23,100	461,000,000	22,100	1,780,000
Hungarian	29,500	296,000,000	26,300	1,130,000
Indonesian	30,400	75,400,000	14,900	1,690,000
Korean	30,800	5,830,000	28,600	821,000
Malayalam	33,100	1,430,000	14,100	328,000
Tamil	2,950	112,000	28,100	842,000
Tagalog	1,530	68,400	12,700	1,090,000
Turkish	29,000	441,000,000	23,700	795,000

Table 2: IARPA Babel vs. Crúbadán type and token frequencies

Language	Babel types	Babel tokens	Crúbadán types	Crúbadán tokens
Guarani	4,920	391,000	3,150	105,000
Georgian	7,550	408,000	33,900	1,190,000
Swahili	5,240	377,000	16,600	1,680,000
Tamil	9,480	521,000	28,100	842,000
Tagalog	5,370	692,000	12,700	1,090,000
Tok Pisin	1,720	479,000	1,520	1,030,000
Turkish	9,170	663,000	23,700	795,000
Zulu	8,610	416,000	26,900	884,000

2.2 Translation to phonemic representation

For each language in the Open Subtitles and IARPA Babel corpora, we assessed whether it would be possible to translate them to phonemic representations. It is difficult to reconstruct stress reliably, so we did not try to capture this information. We successfully created rules that would translate the following languages (corpus name in parentheses, *o* for open substitles, *b*

[2] For instance, the second most frequent word in Open Subtitles for Georgian is з, which (a) does not appear in the Crúbadán Georgian word frequency list and (b) translates to /v/ in Georgian. Therefore, з is not a Georgian word but likely the Russian preposition *в*.

for IARPA Babel): Bulgarian (o), Catalan (o), Greek (o), Georgian (b), Guarani (b), Hungarian (o), Indonesian (o), Korean (o), Malayalam (o), Swahili (b), Tagalog (o, b), Tamil (o, b), Tok Pisin (b), Turkish (o, b), and Zulu (b).

The translation procedure involved creating regular expressions that would match letters to their corresponding segments, conditioned by the context in which they were used, with the most specific context taking precedence over less specific contexts. Finally, sporadic string editing operations were used e.g. to treat gemination as a segment followed by a repetition (e.g. /t,ː/), rather than the same segment repeating twice (e.g. /t,t/). The translation procedures were verified against reference translation words for those languages. The full translation procedure, the translation code, and the rules used to translate each language are all available at https://urielcpublic.s3.amazonaws.com/code/SCiL2020Code-2019-09-15.tbz.

2.3 Calculation of information-theoretic properties

We followed standard practice for calculating the information-theoretic measurements (e.g. Aylett and Turk, 2004; van Son and van Santen, 2005; Bell et al., 2009). We calculated three properties. *Segment frequency* is the unigram probability of each segment in the entire corpus, negative $\log_2$ transformed, ignoring types. *Segment type frequency* is the probability of finding each segment in any word type (negative $\log_2$ transformed). *Segment informativity* (Cohen Priva, 2008, 2015) is the expected value of each segment's surprisal (based on maximum-likelihood estimates), using all the preceding phonemes as context (van Son and Pols, 2003). Peripheral segments are likely to be mis-calculated, as they appear in very few word types. Therefore, we removed all segments that occurred more than 50 times less frequently than the most frequent segment (by token). This step is crucial because many alphabets (e.g. Tamil) provide means to represent sounds that are not part of the basic phonemic inventory of the language. The down side is that some non-peripheral phonemes could also be excluded by this procedure. Had we processed American English (for which our translation procedure could not be used, but which does have pronunciation dictionaries), the exclusion criterion would have only led to the exclusion of /ʒ/ and /ɔɪ/. The exclusion of /ʒ/ would have been legitimate, as it is indeed a peripheral phoneme that occurs in restricted contexts, but /ɔɪ/ is not a peripheral phoneme in American English, it is only infrequent.

We also calculated bigram type and token frequency to estimate whether the environments in which segments are found are comparable. These properties are more sparse, thus they are expected to show more bias across corpora (burstiness and per-genre effects are expected). We used add-one smoothing in order to consider all bigrams across corpora.

2.4 Properties of interest

For all five properties, segment type frequency, segment token frequency, segment informativity, bigram type frequency, and bigram frequency, we compare them across corpora. We calculated Pearson correlations between the estimates in one corpus and the estimates of the same properties in the other corpus. We chose Pearson correlations because the values of the different properties are expected to be consistent across corpora, rather than having the same rank. We also report the median difference in bits for the five properties, as the properties are supposed to be near-identical across corpora, not just correlated.

3 Results

3.1 Segment-level properties

In both corpora, all three properties were highly correlated, as shown in Table 3 for Open Subtitles and Crúbadán, and in Table 4 for IARPA Babel and Crúbadán. Correlations were higher overall between the Open Subtitles corpus and Crúbadán than between the IARPA Babel corpora and Crúbadán. Type frequency correlations

were higher than token frequency correlations, which means that answering questions such as "how many words have that segment" would be less corpus-dependent than asking "how frequent that segment is." Figure 1 illustrates the relationship between segment frequencies across the Open Subtitles and Crúbadán, and Figure 2 illustrates the relationship between segment frequencies across IARPA Babel and Crúbadán. Figures 3 and 4 illustrate the relationship of segment informativity between Open Subtitles and Crúbadán, and between IARPA Babel and Crúbadán, respectively. All four figures show that low correlation is usually centered around specific segments rather than all segments. For instance, Tamil /iː/ is a lot more frequent in Open Subtitles than in Crúbadán. This is likely due to the under-representation of the words நீங்கள் and நீ, /niːnkaɭ/ and /niː/ respectively, both of which are second person pronouns, because they are less frequent in written corpora than in spoken corpora (rank 51 and 36, vs. 3 and 13, respectively). Such discrepancies were more likely to affect segments whose type frequencies were low than segments whose type frequencies were high, as verified in a post-hoc correlation test between the absolute difference between the estimates and their type frequency (always positive, statistically significant in 10 out of the 18 comparisons we have).

Table 3: Open Subtitles vs. Crúbadán correlation between information-theoretic properties. For every property, we provide the Pearson r correlation, and in parentheses, the median absolute difference in bits.

Language	Seg. type freq.	Seg. token freq	Seg. informativity
Bulgarian	0.99 (0.08)	0.97 (0.13)	0.97 (0.17)
Catalan	1 (0.05)	0.99 (0.12)	0.95 (0.24)
Greek	0.99 (0.06)	0.99 (0.16)	0.92 (0.29)
Hungarian	0.99 (0.07)	0.99 (0.14)	0.98 (0.13)
Indonesian	0.99 (0.13)	0.98 (0.19)	0.98 (0.17)
Korean	0.98 (0.13)	0.98 (0.22)	0.96 (0.17)
Malayalam	0.99 (0.1)	0.98 (0.18)	0.99 (0.11)
Tamil	0.98 (0.17)	0.92 (0.19)	0.83 (0.37)
Tagalog	0.98 (0.29)	0.97 (0.11)	0.92 (0.17)
Turkish	0.99 (0.11)	0.99 (0.13)	0.98 (0.14)

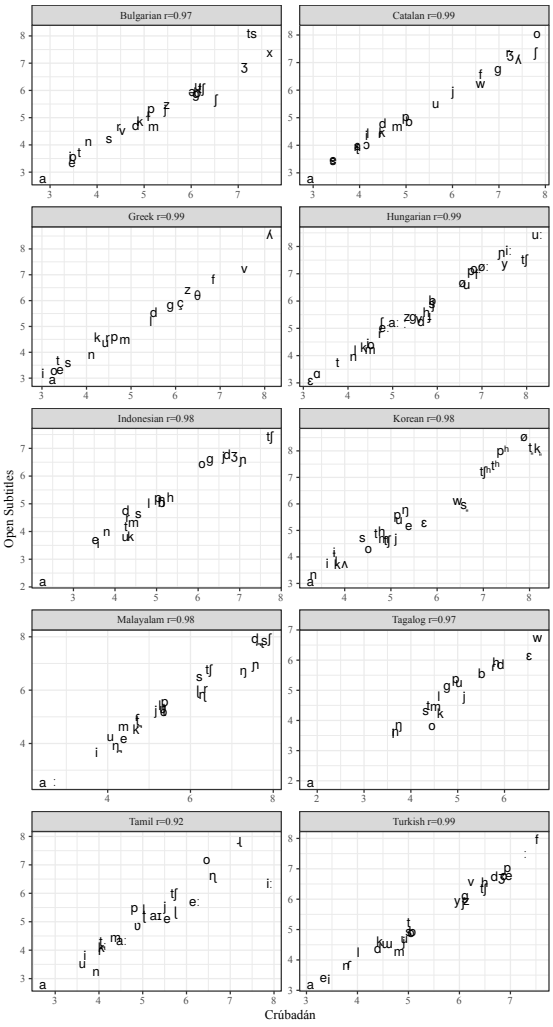

Figure 1: Segment frequency correlation between Open Subtitles and Crúbadán frequency. Both axes are in bits.

3.2 Bigram-level properties

The results are summarized in Table 5 for Open Subtitles and Crúbadán, and in Table 6 for

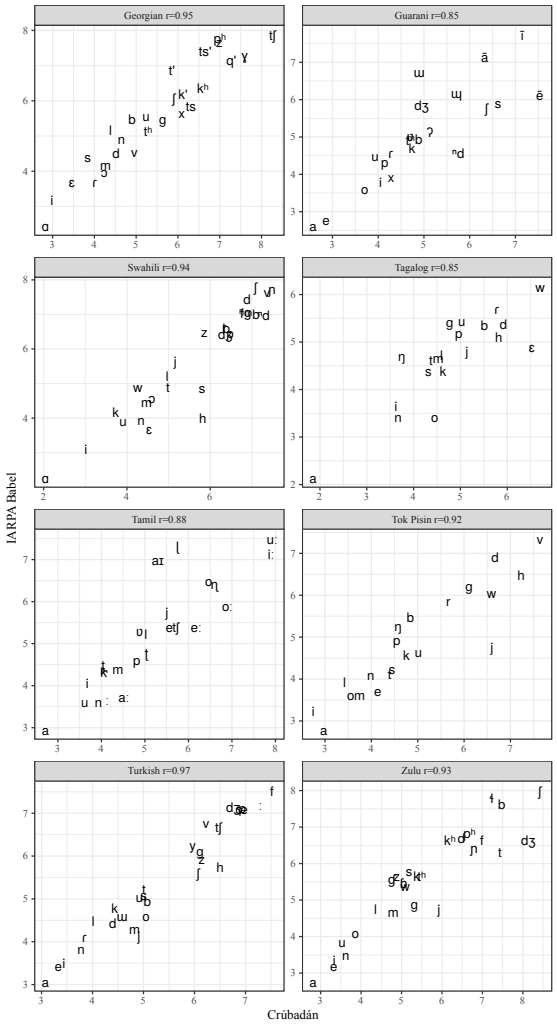

Figure 2: Segment frequency correlation between IARPA Babel and Crúbadán frequency. Both axes are in bits.

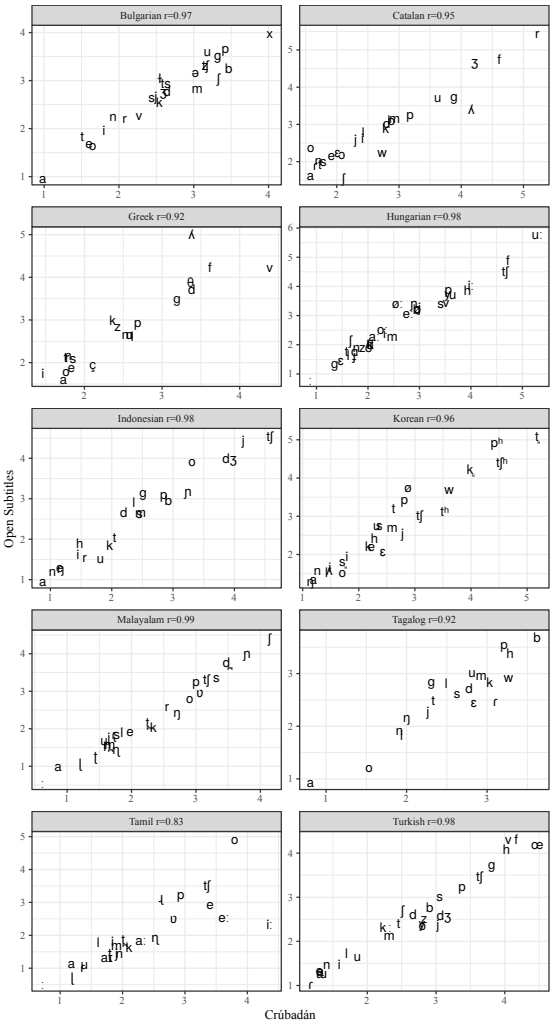

Figure 3: Segment informativity correlation between Open Subtitles and Crúbadán informativity. Both axes are in bits.

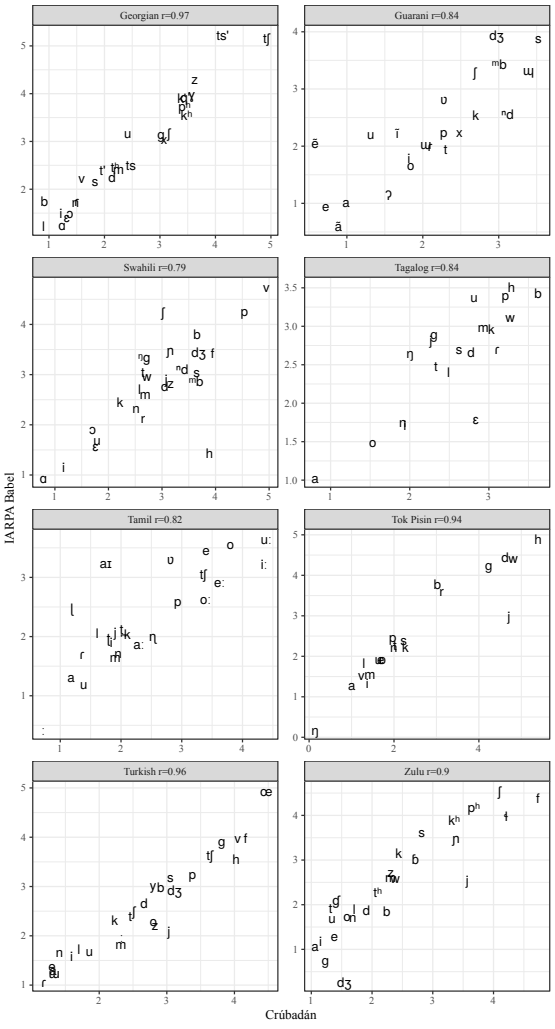

Figure 4: Segment informativity correlation between IARPA Babel and Crúbadán informativity. Both axes are in bits.

Table 4: IARPA Babel vs. Crúbadán correlation between information-theoretic properties. For every property, we provide the Pearson *r* correlation, and in parentheses, the median absolute difference in bits.

Language	Seg. type freq.	Seg. token freq	Seg. informativity
Guarani	0.9 (0.19)	0.85 (0.35)	0.84 (0.32)
Georgian	0.98 (0.22)	0.95 (0.32)	0.97 (0.27)
Swahili	0.96 (0.24)	0.94 (0.27)	0.79 (0.26)
Tamil	0.94 (0.3)	0.88 (0.33)	0.82 (0.3)
Tagalog	0.95 (0.17)	0.85 (0.27)	0.84 (0.19)
Tok Pisin	0.95 (0.22)	0.92 (0.3)	0.94 (0.25)
Turkish	0.99 (0.15)	0.97 (0.18)	0.96 (0.12)
Zulu	0.95 (0.32)	0.93 (0.36)	0.9 (0.34)

IARPA Babel and Crúbadan.

As expected, the correlations were overall lower at the bigram level than at the segmental level, likely due to sparsity issues that we know exist at the word level (Daland, 2013). However, for most languages, the correlations were still impressively high, at Pearson r>.93 and r>.85 for bigram type frequency, representative of Open Subtitles and IARPA Babel's correlations with Crúbadán respectively, and Pearson r>.86 and r>.79 for bigram token frequencies, representative of Open Subtitles and IARPA Babel's correlations with Crúbadán respectively. For reference, assuming that the inherent noise of an experimental population is SD=1 and the sampling noise equals SD=.5, the correlation between test and retest of the same individual is expected to be around Pearson r=.8.

Table 5: Open Subtitles vs. Crúbadán correlation between type and token frequencies for bigrams. For every property, we provide the Pearson *r* correlation, and in parentheses, the median absolute difference in bits.

Language	# bigram types	Bigram type freq.	Bigram token freq
Bulgarian	608	0.98 (0.27)	0.86 (0.56)
Catalan	611	0.97 (0.28)	0.94 (0.58)
Greek	464	0.97 (0.33)	0.88 (0.63)
Hungarian	1202	0.96 (0.39)	0.83 (0.6)
Indonesian	627	0.97 (0.36)	0.91 (0.74)
Korean	705	0.96 (0.45)	0.9 (0.67)
Malayalam	970	0.95 (0.4)	0.9 (0.71)
Tamil	681	0.94 (0.62)	0.9 (0.87)
Tagalog	446	0.93 (0.63)	0.89 (0.74)
Turkish	733	0.97 (0.41)	0.85 (0.63)

Table 6: IARPA Babel vs. Crúbadán correlation between type and token frequencies for bigrams. For every property, we provide the Pearson r correlation, and in parentheses, the median absolute difference in bits.

Language	# bigram types	Bigram type freq.	Bigram token freq
Guarani	484	0.85 (0.91)	0.74 (1.68)
Georgian	879	0.93 (0.66)	0.88 (1.29)
Swahili	621	0.91 (0.74)	0.81 (1.12)
Tamil	714	0.9 (0.91)	0.81 (1.47)
Tagalog	479	0.91 (0.48)	0.79 (1.39)
Tok Pisin	357	0.9 (0.56)	0.83 (1.14)
Turkish	724	0.96 (0.43)	0.91 (1.22)
Zulu	729	0.87 (0.81)	0.73 (1.58)

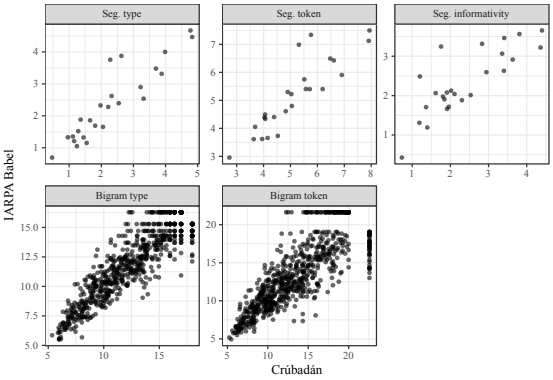

Figure 5: Segment type frequency, token frequency, and informativity, as well as bigram type frequency and bigram informativity for Tamil, by property. Especially for bigram values, it is evident that estimates get progressively worse for low frequency values.

4 Discussion

4.1 Differences across corpora and corpus-usability

We were concerned that the lower correlations between IARPA Babel and Crúbadán, relative to the correlations between Open Subtitles and Crúbadan, were due to the smaller size of the corpus. Cohen Priva and Jaeger (2018) report correlations that approximate >.99 for segment frequency with as few as 100,000 word tokens, a threshold nearly all of our corpora passed (except Open Subtitles for Tagalog). To verify that corpus size is not an issue we ran a post-hoc analysis to predict segment correlations (Fisher-transformed) using log frequencies from the two contributing corpora. Except for a marginal effect for token frequencies in Open Subtitles, there was no correlation. We did observe substantially more interjections, false-starts, loanwords, and conversation-starting / ending in IARPA Babel than in either Crúbadán or Open Subtitles, which is to be expected given the type of the corpus. We are not sure why different languages show this effect to different extents, but given the number of comparisons we have, it would seem that the lower boundary on within-language correlations is still high enough to support the study of phonological properties using corpora of different types and with relatively high degrees of noise.

4.2 Reducing noise

Given that some degree of noise does exist when switching corpus types, it is important to ask what could be done to decrease the amount of noise. One parameter researchers can control is reliance on low-frequency segments and bigrams as well as the use of more robust statistics.

Certainty of information-theoretic values diminishes for less frequent segments and bigrams, which are more easily swayed by word-level frequency effects. Figure 5 shows the correlations for Tamil. It is evident that the estimates for lower-frequency bigrams (and to some extent, individual segments) are worse than for high-frequency segments. Studies that cannot tolerate the lower-precision that is associated with changes across genres could therefore focus on high-frequency segments and contexts.

5 Conclusion

We checked whether segment type frequency, segment token frequency, segment informativity, as well as bigram type and token frequency could be reliably estimated across different corpus types genres. We showed that segments were more reliably estimated than bigrams and that type frequencies were more reliably estimated than token frequencies. However, even for the least similar corpora, Crúbadán and

IARPA Babel, the reliability of measurements was substantial, and likely not larger than for many experimental designs. We also found that high-frequency elements were more reliably estimated than lower-frequency ones. We therefore believe that corpus-based research in phonology can mitigate the concerns related to generalizations across genre and corpus types.

6 Acknowledgements

The work presented in this paper was supported by NSF Grant, awarded to the first author, BCS-1829290. Many thanks go out to Justin Bai (Indonesian, Korean, Tok Pisin, Turkish), Abi Creighton (Catalan, Tagalog, Zulu), Maddie Critz (Georgian), Delphine Morse Mahos (Swahili), Becky Mathew (Indonesian, Korean, Malayalam, Georgian, Tamil), and Bill Mizgerd (Bulgarian, Guarani) for creating the transcription rules necessary for this research to be conducted. The languages each individual was responsible for are indicated in parentheses. We would also like to thank Robert Daland and the two anonymous reviewers for providing constructive feedback on this research.

References

Nikki Adams, Aric Bills, Thomas Conners, Eyal Dubinski, Jonathan G Fiscus, Mary Harper, Willa Lin, Jennifer Melot, Jessica Ray, Anton Rytting, Wade Shen, Ronnie Silber, Evelyne Tzoukermann, and Jamie Wong. 2017. IARPA Babel Zulu Language Pack IARPA-babel206b-v0.1e LDC2017s19. Technical report, Linguistic Data Consortium, Philadelphia, PA.

Jess Andresen, Aric Bills, Thomas Conners, Eyal Dubinski, Jonathan G Fiscus, Mary Harper, Kirill Kozlov, Nicolas Malyska, Jennifer Melot, Michelle Morrison, Josh Phillips, Jessica Ray, Anton Rytting, Wade Shen, Ronnie Silber, Evelyne Tzoukermann, and Jamie Wong. 2017. IARPA Babel Swahili Language Pack IARPA-babel202b-v1.0d. LDC2017s05. Technical report, Linguistic Data Consortium, Philadelphia, PA.

Lucy Andresen, Aric Bills, Claudia Brugman, Thomas Conners, Anne David, Eyal Dubinski, Jonathan G Fiscus, Ketty Gann, Mary Harper, Michael Kazi, Hanh Le, Nicolas Malyska, Arlene Maurillo, Jennifer Melot, Shelley Paget, Jane Elizabeth Prebble, Jessica Ray, Fred Richardson, Anton Rytting, and Sinney Shen. 2019. IARPA Babel Guarani language pack IARPA-babel305b-v1.0c LDC2019s08. Technical report, Linguistic Data Consortium, Philadelphia, PA.

Lucy Andresen, Aric Bills, Thomas Conners, Luanne Dela Cruz, Eyal Dubinski, Jonathan G Fiscus, Mary Harper, Hanh Le, Arlene Maurillo, Jennifer Melot, Josh Phillips, Jessica Ray, Anton Rytting, Wade Shen, Ronnie Silber, and Evelyne Tzoukermann. 2018. IARPA Babel Cebuano Language Pack IARPA-babel301b-v2.0b LDC2018s07. Technical report, Linguistic Data Consortium, Philadelphia, PA.

Tony Andrus, Aric Bills, Thomas Conners, Erin Smith Crabb, Eyal Dubinski, Jonathan G Fiscus, Breanna Gillies, Mary Harper, T.J Hazen, Brook Hefright, Amy Jarrett, Hanh Le, Jessica Ray, Anton Rytting, Wade Shen, Ronnie Silber, and Evelyne Tzoukermann. 2017a. IARPA Babel Haitian Creole Language Pack IARPA-babel201b-v0.2b LDC2017s03. Technical report, Linguistic Data Consortium, Philadelphia, PA.

Tony Andrus, Aric Bills, Miriam Corris, Eyal Dubinski, Jonathan G Fiscus, Breanna Gillies, Mary Harper, T.J Hazen, Brook Hefright, Amy Jarrett, Hanh Le, Jessica Ray, Anton Rytting, Ronnie Silber, Wade Shen, and Evelyne Tzoukermann. 2017b. IARPA Babel Vietnamese Language Pack IARPA-babel107b-v0.7. LDC2017s01. Technical report, Linguistic Data Consortium, Philadelphia, PA.

Arto Anttila. 1997. Deriving variation from grammar: a study of Finnish genitives. In Frans Hinskens, Roeland van Hout, and Leo Wetzels, editors, *Variation, change and phonological theory*, pages 35–68. John Benjamins, Amsterdam.

Matthew Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56.

Alan Bell, Jason Brenier, Michelle Gregory, Cynthia Girand, and Daniel Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1):92–111.

Daniel Benowitz, Aric Bills, Thomas Conners, Anne David, Eyal Dubinski, Jonathan G Fiscus, Mary Harper, Brook Hefright, Hanh Le, Jennifer Melot, Jessica Ray, Anton Rytting, Wade Shen, Rosanna Smith, Wade Shen, Ronnie Silber, and Evelyne Tzoukermann. 2019. IARPA Babel Lithuanian Language Pack IARPA-babel304b-v1.0b. LDC2019s03. Technical report, Linguistic Data Consortium, Philadelphia, PA.

Aric Bills, Thomas Conners, Miriam Corris, Anne David, Eyal Dubinski, Jonathan G Fiscus, Mary Harper, Alice Kaiser-Schatzlein, Jennifer Melot, Shelley Paget, Jessica Ray, Anton Rytting, Wade Shen, Ronnie Silber, Evelyne Tzoukermann, and Arun Viswanath. 2015. IARPA Babel Tamil Language Pack IARPA-babel204b-v1.1b. LDC2017s13. Technical report, Linguistic Data Consortium, Philadelphia, PA.

Aric Bills, Thomas Conners, Anne David, Eyal Dubinski, Jonathan G Fiscus, Mary Harper, Brook Hefright, Kirill Kozlov, Jennifer Melot, Jessica Ray, Anton Rytting, Josh Phillips, Marle Walter, Wade Shen, Ronnie Silber, and Evelyne Tzoukermann. 2018. IARPA Babel Kazakh Language Pack IARPA-babel302b-v1.0a. LDC2018s13. Technical report, Linguistic Data Consortium, Philadelphia, PA.

Aric Bills, Anne David, Eyal Dubinski, Jonathan G Fiscus, Simon Hammond, Ketty Gann, Mary Harper, Brook Hefright, Michael Kazi, Julie Lam, Jessica Ray, Fred Richardson, Anton Rytting, and Marle Walter. 2016. IARPA Babel Georgian Language Pack IARPA-babel404b-v1.0a LDC2016s12. Technical report, Linguistic Data Consortium, Philadelphia, PA.

Paul Boersma. 1998. *Functional Phonology*. Ph.D. thesis, University of Amsterdam.

Uriel Cohen Priva. 2008. Using information content to predict phone deletion. In *Proceedings of the 27th West Coast Conference on Formal Linguistics*, pages 90–98, Somerville, MA. Cascadilla Proceedings Project.

Uriel Cohen Priva. 2015. Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, 6(2):243–278.

Uriel Cohen Priva. 2017. Informativity and the actuation of lenition. *Language*, 93(3):569–597. (preprint).

Uriel Cohen Priva and T. Florian Jaeger. 2018. The interdependence of frequency, predictability, and informativity. *Linguistics Vanguard*, 4.

Thomas Conners, Jonathan G. Fiscus, Breanna Gillies, Mary Harper, T. J. Hazen, Amy Jarrett, Willa Lin, María Encarnación Pérez Molina, Shawna Rafalko, Jessica Ray, Anton Rytting, Wade Shen, and Evelyne Tzoukermann. 2016. IARPA Babel Tagalog Language Pack IARPA-babel106-v0.2g. LDC2016s13. Technical report, Linguistic Data Consortium, Philadelphia, PA.

Robert Daland. 2013. Variation in the input: a case study of manner class frequencies. *Journal of Child Language*, 40:1091–1122.

Sharon Goldwater and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm workshop on variation within Optimality Theory*, pages 111–120.

John J. McCarthy. 1986. OCP effects: Gemination and antigemination. *Linguistic Inquiry*, 17(2):207–263.

Steven T. Piantadosi, Harry J Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*.

Kevin P Scannell. 2007. The Crúbadán project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 4, pages 5–15.

R. J. J. H. van Son and L. C. W. Pols. 2003. How efficient is speech? *Proceedings of the Institute of Phonetic Sciences*, 25:171–184.

R.J.J.H. van Son and J.P.H van Santen. 2005. Duration and spectral balance of intervocalic consonants: a case for efficient communication. *Speech Communication*, 47:100–123.

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Kie Zuraw. 2006. Using the web as a phonological corpus: A case study from Tagalog. In *Proceedings of the 2Nd International Workshop on Web As Corpus*, WAC '06, pages 59–66, Stroudsburg, PA, USA. Association for Computational Linguistics.